A tool for sharing interlinearized and lexical data in diverse formats

Daniel Kaufman  Raphael Finkel
(ELA & Queens College) (University of Kentucky)
Quick demo

- Creating an account
- Searching for a word in Kratylos
- Searching across files/projects
- Exporting an example
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - Various current and legacy formats for media annotation, interlinearized glossed text and lexicons

Arkut & Finkel
Kratylos
March 2, 2017
3 / 48
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - Various current and legacy formats for media annotation, interlinearized glossed text and lexicons
- Typically, there is electronic data from multiple scholars on a single language or a group of closely related languages
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - Various current and legacy formats for media annotation, interlinearized glossed text and lexicons
- Typically, there is electronic data from multiple scholars on a single language or a group of closely related languages
- How can this data be tamed? How can other scholars and community members make use of it?
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - Various current and legacy formats for media annotation, interlinearized glossed text and lexicons
- Typically, there is electronic data from multiple scholars on a single language or a group of closely related languages
- How can this data be tamed? How can other scholars and community members make use of it?
- Archives have not yet provided a solution. (They have enough problems of their own.)
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - Various current and legacy formats for media annotation, interlinearized glossed text and lexicons
- Typically, there is electronic data from multiple scholars on a single language or a group of closely related languages
- How can this data be tamed? How can other scholars and community members make use of it?
- Archives have not yet provided a solution. (They have enough problems of their own.)
- There is no widely available system for eliciting feedback and corrections on linguistic data from a community of speakers
The problem

- Long-term language documentation projects accrue data in different electronic formats.
  - Audio files and video files
  - various current and legacy formats for media annotation, interlinearized glossed text and lexicons
- Typically, there is electronic data from multiple scholars on a single language or a group of closely related languages
- How can this data be tamed? How can other scholars and community members make use of it?
- Archives have not yet provided a solution. (They have enough problems of their own.)
- There is no widely available system for eliciting feedback and corrections on linguistic data from a community of speakers
- Existing electronic lexicons are still very much based on print models despite having none of the constraints of traditional print dictionaries.
Our solution in progress

- **Kratylos**: a website that can store and search over FLEX databases, Shoebox, Toolbox, Praat and ELAN files including associated audio/video media.
Our solution in progress

- **Kratylos**: a website that can store and search over FLEx databases, Shoebox, Toolbox, Praat and ELAN files including associated audio/video media
- While it is **not** an archiving facility, it can salvage data trapped in formats that are otherwise difficult to share.
Our solution in progress

- **Kratylos**: a website that can store and search over FLEex databases, Shoebox, Toolbox, Praat and ELAN files including associated audio/video media.
- While it is **not** an archiving facility, it can salvage data trapped in formats that are otherwise difficult to share.
- We demonstrate this with the Wakhi language documentation project.
We are not aiming to replace popular existing software like FLEn. We are aiming to make data from such programs more easily sharable.

A bit more about FLEn:
- FLEn builds on Shoebox and Toolbox, two older and far simpler pieces of software for creating linguistic databases.
- It has wide-ranging features for building sophisticated lexicons + a very detailed approach to interlinearized glossed text.
5.5 Word

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>jaw</th>
<th>δaj</th>
<th>wazdaj</th>
<th>tam</th>
<th>prid</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex. Entries</td>
<td>jaw</td>
<td>δaj</td>
<td>woz</td>
<td>-d</td>
<td>-aj</td>
</tr>
<tr>
<td>Lex. Gloss</td>
<td>DEF/3SG</td>
<td>man</td>
<td>PST</td>
<td>PST</td>
<td>LOC.UP</td>
</tr>
<tr>
<td>Lex. Gram. Info.</td>
<td>det</td>
<td>n</td>
<td>v</td>
<td>v:Past1</td>
<td>v:(Past2)</td>
</tr>
<tr>
<td>Word Gloss</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Word Cat.</td>
<td>pro</td>
<td>n</td>
<td>***</td>
<td>***</td>
<td>***</td>
</tr>
</tbody>
</table>

Free: If he says "Why dost thou not come forth?" say: "Thou dost not love me; if thou lovest me give me (thy) ring."

5.6 Word

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>čataj</th>
<th>t</th>
<th>ki</th>
<th>&quot;</th>
<th>třizarat</th>
<th>zj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex. Entries</td>
<td>čan</td>
<td>-t</td>
<td>-j</td>
<td>ki</td>
<td>třiz</td>
<td>-ar</td>
</tr>
<tr>
<td>Lex. Gloss</td>
<td>say</td>
<td>PST</td>
<td>PST</td>
<td>COMP</td>
<td>what</td>
<td>DAT</td>
</tr>
<tr>
<td>Lex. Gram. Info.</td>
<td>v</td>
<td>v:Past1</td>
<td>v:(Past2)</td>
<td>comp</td>
<td>interrog</td>
<td>n/(Outer case)</td>
</tr>
<tr>
<td>Word Gloss</td>
<td>said</td>
<td>v</td>
<td>comp</td>
<td>***</td>
<td>***</td>
<td></td>
</tr>
<tr>
<td>Word Cat.</td>
<td>v</td>
<td>comp</td>
<td>***</td>
<td>***</td>
<td>***</td>
<td></td>
</tr>
</tbody>
</table>

Free: Her husband came, she went not forth into his presence.

5.7 Word

<table>
<thead>
<tr>
<th>Morphemes</th>
<th>δajč</th>
<th>čataj</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lex. Entries</td>
<td>δaj</td>
<td>čan</td>
</tr>
<tr>
<td>Lex. Gloss</td>
<td>daughter</td>
<td>say</td>
</tr>
<tr>
<td>Lex. Gram. Info.</td>
<td>n</td>
<td>v</td>
</tr>
<tr>
<td>5.5 Word</td>
<td>jaw</td>
<td>ḍaj</td>
</tr>
<tr>
<td>----------</td>
<td>-----</td>
<td>-----</td>
</tr>
<tr>
<td>Morphemes</td>
<td>jaw</td>
<td>ḍaj</td>
</tr>
<tr>
<td>Lex. Entries</td>
<td>ja₂</td>
<td>ဗ</td>
</tr>
<tr>
<td>Lex. Gloss</td>
<td>DEF/3SG</td>
<td>man</td>
</tr>
<tr>
<td>Lex. Gram. Info.</td>
<td>det</td>
<td>n</td>
</tr>
<tr>
<td>Word Gloss</td>
<td>***</td>
<td>***</td>
</tr>
<tr>
<td>Word Cat.</td>
<td>pro</td>
<td>n</td>
</tr>
</tbody>
</table>

*Free* Her husband came, she went not forth into his presence.
5.5 Word

| Morphemes | jaw | δaj | wazδaj |
| Lex. Entries | jaw | δaj | waz | -d | -əj |
| Lex. Gloss | ja₂ | δaj | waz | -t₁ | -i₁ |
| Loc./Dir | LOC.UP | PROX | Loc./Dir | det | (Proximity) | n | <Not Sure> | v |
| Word Gloss | det | n | v | v:Past1 | v:(Pst2) |
| Word Cat. | *** | *** | *** |

Free: Her husband came, she went not forth into his presence.
| Morpahmes | jaw | δaj | wazđaj | -d | -əj |
| Lex. Entries | ja₂ | δaj | waz | -t₂ | -i₁ |
| Lex. Gloss | DEF/3SG | man | come | PST | PST |
| Lex. Gram. Info. | det | n | v | v:Past1 | v:(Pst2) |
| Word Gloss | *** | *** | *** | *** | *** |
| Word Cat. | pro | n | *** | Loc/Dir | n | *** | v |

**Free** Her husband came, she went not forth into his presence.
wazdeaj

waz   -d     -eaj
waz   -t₂     -i₁
come   PST     PST
v:Past₁   v:(Pst₂)

***

***
The big lacuna (recap)

- FLEx projects are largely stuck in FLEx - no means of sharing database functions.
- More generally, little means for complex searches over interlinearized text on the web.
- Multiple researchers may have data sets on the same language that are difficult to unify.
- For existing lexicon web apps (e.g. Webonary, Lexique Pro), no possibility for regular expression searches or for searching over multiple languages simultaneously.
An attempt at a solution: Kratylos

- Supported by NSF DEL Grant #1500753, which includes database work and fieldwork on Wakhi (Iranic; Pakistan, Tajikistan and Afghanistan), Purhepecha (language isolate; Mexico) and Koda (Munda; Bangladesh).
- Programming work: Raphael Finkel and RA Jiho Noh at the University of Kentucky.
Behind the scenes

- The web server, Apache2, invokes Perl scripts using the Common Gateway Interface (CGI). The scripts use several standard modules:
  - CGI (and submodules Carp, Session, and cookie), HTML::Template, Digest (submodules SHA and MD5), JSON, and Unicode::Normalize.

- Kratylos converts uploaded data, if necessary, into a new XML format. For example, the EAF format, although in XML, is not divided into entries, so Kratylos reformats it into entries, each of which contains all the relevant tiers (such as headword, part of speech, and gloss) and a reference to the media file.
Behind the scenes

- Kratylos then applies a template to convert the XML into a Qdcb (Quick and dirty database) representation.
- The template is format-specific and coordinates:
  1. the XML fields, described as XPath expressions
  2. the Qdcb representation of those fields, which is hierarchical
  3. the formatting that the linear display should employ for those fields, which involves Cascading Style Sheets (CSS).

- The web pages that Kratylos presents to the user use the Bootstrap and Jquery libraries to format pages. The query results page also contains JavaScript code that converts entries on the fly text.
- Kratylos maintains a MySQL database coordinating projects with their owners and other information.
Sharing versus archiving

- Optimally, such a tool should exist within an archive so you have complex searches over a stable set of data.
Behind the scenes

Sharing versus archiving

▶ Optimally, such a tool should exist within an archive so you have complex searches over a stable set of data.
▶ But we are approaching the problem from the linguist’s perspective rather than the archivist’s perspective.
Sharing versus archiving

- Optimally, such a tool should exist within an archive so you have complex searches over a stable set of data.
- But we are approaching the problem from the linguist’s perspective rather than the archivist’s perspective.
- And hoping to help bridge the gap between archiving and publication.
  - Citations are important but absolute permanence and completely standard metadata for each piece of data is beyond our reach.
Behind the scenes

Kratylos ([http://www.kratylos.org](http://www.kratylos.org))

- Table of public files which can be searched over. Logging in reveals your private files.

<table>
<thead>
<tr>
<th>Language</th>
<th>Version</th>
<th>Data Sources</th>
<th>Access</th>
<th>Maintainer</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aceh-Chamic</td>
<td>base</td>
<td>Parallel Bible Corpus (paralleltext.info)</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Amuzgoan</td>
<td>base</td>
<td>Amy Bauernschmidt (2013)</td>
<td>public</td>
<td>Jiho Noh (University of Kentucky)</td>
</tr>
<tr>
<td>Aramaic</td>
<td>aramaic2</td>
<td>Daniel Kaufman, 28 December 2015</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Aramaic</td>
<td>aramaic1</td>
<td>Daniel Kaufman, 28 December 2015</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Archi</td>
<td>archi1</td>
<td>Marina Chumakina</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Balantak</td>
<td>base</td>
<td>Parallel Bible Corpus (paralleltext.info)</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Bambam</td>
<td>base</td>
<td>Parallel Bible Corpus (paralleltext.info)</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Batak</td>
<td>base</td>
<td>Parallel Bible Corpus (paralleltext.info)</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Bugis</td>
<td>base</td>
<td>Parallel Bible Corpus (paralleltext.info)</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
<tr>
<td>Bukvar</td>
<td>base</td>
<td>Elena Pererhvalskaya, 2/2016</td>
<td>public</td>
<td>Raphael Finkel (University of Kentucky)</td>
</tr>
</tbody>
</table>

Showing 1 to 63 of 63 entries
show project list

Selected Languages:

Select languages above then search a word or a phrase...
Behind the scenes

Kratylos (http://www.kratylos.org)

- A search bar which can be set to query full words, strings or regular expressions ("patterns").
- Any data field can be targeted by a query. The options are populated automatically by the categories in the data.
Krtylos (http://www.krtylos.org)

- Displaying particular fields can be toggled on and off for different audiences.
Linear display uses formatting to distinguish fields.

**gurung (lift)**  Data source: Recorded from Narayan Gurung of Siklis village in the Endangered Language Initiative at CUNY Graduate Center. Researchers: Peter Graif,

2014-03-24T14:41:03Z
2014-03-24T14:41:03Z

gvr
*tíli kʰora:
stem
en
**hog house**


Outline display uses text to distinguish fields.

<table>
<thead>
<tr>
<th>DataCreated</th>
<th>2014-03-24T14:41:03Z</th>
</tr>
</thead>
<tbody>
<tr>
<td>DataModified</td>
<td>2014-03-24T14:41:03Z</td>
</tr>
<tr>
<td>Headword</td>
<td>gvr</td>
</tr>
<tr>
<td>HLanguage</td>
<td>gvr</td>
</tr>
<tr>
<td>HText</td>
<td><em>tíli kʰora:</em></td>
</tr>
<tr>
<td>MorphologicalType</td>
<td>stem</td>
</tr>
<tr>
<td>Sense</td>
<td></td>
</tr>
<tr>
<td>Gloss</td>
<td>en</td>
</tr>
<tr>
<td>GLanguage</td>
<td>en</td>
</tr>
<tr>
<td><strong>GlossData</strong></td>
<td><strong>hog house</strong></td>
</tr>
</tbody>
</table>
Wakhi project

- An Iranian language spoken by a small transnational community around the intersection of Afghanistan, Tajikistan, Pakistan, and China.
The relation between Wakhi and the other Iranian languages is still unclear but it is often grouped together with the Pamiri languages (Shughni, Roshani, Bartangi, Oroshori, Sarikoli, Munji, Ishkashimi). Recent work suggests the Pamiri languages constitute a sprachbund rather than a phylogenetic group.
Behind the scenes

Indo-European
  | Indo-Iranian
  | Indo-Aryan Iranian
    | Western
    | Eastern
      | Persian
      | Pashto
      | Pamir
        | Ishkashimi
        | Munji
        | Sanglechi
        | Yidgha
        | Shugni-Yazgulami
          | Sarikoli
          | Shughni
          | Yazgulyam
Beginnings of the project: A small immigrant community in NYC
Behind the scenes

Beginnings of the project: From NYC to the Wakhan

kuj jaw dijt jaw mərz jo ṭax
Who knows whether he is thirsty or hungry?
Previous literature


For our purposes, Pakhalina (1975) and Grünberg and Steblin-Kamensky (1988) are the most crucial because of the folktales, narratives, and songs they recorded. These are inaccessible to the Wakhis of Afghanistan and Pakistan because they are written in Russian.
Previous literature


- For our purposes, Pakhalina (1975) and Grünberg and Steblin-Kamensky (1988) are the most crucial because of the folktales, narratives and songs they recorded. These are inaccessible to the Wakhis of Afghanistan and Pakistan because they are written in Russian.
Wakhi writing systems
# Wakhi writing systems

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>Ć</th>
<th>Ć</th>
<th>D</th>
<th>D̞</th>
<th>Δ</th>
<th>Δ̞</th>
<th>E</th>
<th>E̞</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>b</td>
<td>ц</td>
<td>ч</td>
<td>ч</td>
<td>d</td>
<td>d̞</td>
<td>d̞</td>
<td>d̞</td>
<td>e</td>
<td>e̞</td>
</tr>
<tr>
<td>[a]</td>
<td>[b]</td>
<td>[ц]</td>
<td>[ч]</td>
<td>[ч]</td>
<td>[d]</td>
<td>[d̞]</td>
<td>[d̞]</td>
<td>[d̞]</td>
<td>[e]</td>
<td>[e̞]</td>
</tr>
<tr>
<td>F</td>
<td>G</td>
<td>Ы</td>
<td>Ы</td>
<td>H</td>
<td>I</td>
<td>J</td>
<td>Ј</td>
<td>K</td>
<td>L</td>
<td></td>
</tr>
<tr>
<td>Ф</td>
<td>Г</td>
<td>Ы</td>
<td>Ы</td>
<td>Х</td>
<td>И</td>
<td>Ч</td>
<td>Ч</td>
<td>Ч</td>
<td>К</td>
<td></td>
</tr>
<tr>
<td>l</td>
<td>g</td>
<td>y</td>
<td>y</td>
<td>h</td>
<td>i</td>
<td>j</td>
<td>j</td>
<td>k</td>
<td>l</td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>m</td>
<td>O</td>
<td>O</td>
<td>P</td>
<td>P</td>
<td>Q</td>
<td>Q</td>
<td>R</td>
<td>r</td>
<td></td>
</tr>
<tr>
<td>М</td>
<td>м</td>
<td>О</td>
<td>О</td>
<td>П</td>
<td>П</td>
<td>С</td>
<td>С</td>
<td>Ш</td>
<td>Ш</td>
<td></td>
</tr>
<tr>
<td>t</td>
<td>t</td>
<td>ё</td>
<td>ё</td>
<td>u</td>
<td>u</td>
<td>v</td>
<td>v</td>
<td>w</td>
<td>w</td>
<td></td>
</tr>
<tr>
<td>Т</td>
<td>т</td>
<td>ё</td>
<td>ё</td>
<td>у</td>
<td>у</td>
<td>х</td>
<td>х</td>
<td>х̆</td>
<td>х̆</td>
<td></td>
</tr>
<tr>
<td>z</td>
<td>z</td>
<td>й</td>
<td>й</td>
<td>Z</td>
<td>Z</td>
<td>З</td>
<td>З</td>
<td>Ж</td>
<td>Ж</td>
<td></td>
</tr>
<tr>
<td>[t]</td>
<td>[т]</td>
<td>[у]</td>
<td>[у]</td>
<td>[w]</td>
<td>[w]</td>
<td>[x]</td>
<td>[x]</td>
<td>[j]</td>
<td>[z]</td>
<td></td>
</tr>
<tr>
<td>ž</td>
<td>ž</td>
<td>з</td>
<td>з</td>
<td>Ы</td>
<td>Ы</td>
<td>Ж</td>
<td>Ж</td>
<td>Ч</td>
<td>Ч</td>
<td></td>
</tr>
</tbody>
</table>

**Kaufman & Finkel**

**Krtylos**

March 2, 2017 24/48
Wakhi writing systems
Wakhi writing systems


Xhik Murdumver slom; Ghafch bef yark oghozve xhetk, zu ghena saver muboraki. Zunini her yi xhik murdumer yi arz woz swol, cum batkenep sak sach vir re samander mulung wozen din woz dak vanden ki koipe wizit xe dhastep saker rashufd, wazet oghoz ceren, terleman wocn. Wodh ki nei peghrep sak xath xath te xhu dhasten nusun. her yiu shenoxti yo zik, wazeth xhath ce neshaken shetraxven.

Salam to Wakhi people around the globe, very niece effort is going on, many congratulation to all of you. I have a a request and quest from each Wakhi people. For how long we will swim alone in the Ocean and waite for others to lend their hand to us for help? Lets come together and and start a sincere effort. If not today tomorrow we will kill ourselves with our own hands. Language is identity of any individual, lets come forward and
Exploring Wakhi morphosyntax with regular expressions

- Fuchs (2015) shows that certain clitics in Wakhi can be doubled and tripled within a single clause:

1) \(\text{wuz} = \text{əm} \quad \text{çi} \quad \text{ʃew-}i = \text{m} \quad \text{ʃkəndevdi}\)

1SG.NOM = 1SG SELF.POSS horns-ACC = 1SG break.PST

‘I broke my horns.’ (Fuchs 2015:151)

2) \(\text{jan} \quad \text{də} \quad \text{sɔbɪq} = \text{əʂ} \quad \text{lup} = \text{əʂ} \quad \text{çat-i} \quad \text{ki} \quad \text{nɔyardum} \quad \text{ətf} \quad \text{waxt} \quad \text{də} \quad \text{then in old.time=IPFV big=IPFV say-PST COMP bear} \quad \text{no time with}

\text{mɒltuq-ən} \quad \text{ja-} \quad \text{qrib} \quad \text{me} \quad \text{rəʈʂ-əv}

\text{shotgun-ABL 3SG-DAT near NEG.IMP go-2PL.SAGR}

‘Then in old times the old people said: Bears, never go near them (not even) with a shotgun.’ (SanGregory 2015:7).
Exploring Wakhi morphosyntax with regular expressions

- Fuchs (2015) shows that certain clitics in Wakhi can be doubled and tripled within a single clause:
- She also shows that the positioning of the clitics may depend on focus and other semantic factors (see also, Erschler 2010; Beck 2013; Hughes 2011; Bashir 1986).

(1)  
\[wuz = \text{əm} \quad \text{çi} \quad \text{ʂew-i = m} \quad \text{ʂkəndevdi}\]  
1SG.NOM=1SG SELF.Poss horns-ACC=1SG break.PST  
‘I broke my horns.’ (Fuchs 2015:151)

(2)  
\[\text{jan} \quad \text{də} \quad \text{səbıp = e} \quad \text{lup = e} \quad \text{çat-i} \quad \text{ki} \quad \text{nəɣərdum} \quad \text{əʃ f wuxt də then in old.time=IPFV big=IPFV say-PST COMP bear}\]  
\[\text{nətə} \quad \text{mə} \quad \text{rəʈʂ-əv} \quad \text{shotgun-ABL 3SG-DAT near NEG.IMP go-2PL.SAGR}\]  
‘Then in old times the old people said: Bears, never go near them (not even) with a shotgun.’ (SanGregory 2015:7).
Exploring Wakhi morphosyntax with regular expressions

- Fuchs (2015) shows that certain clitics in Wakhi can be doubled and tripled within a single clause:

- She also shows that the positioning of the clitics may depend on focus and other semantic factors (see also, Erschler 2010; Beck 2013; Hughes 2011; Bashir 1986).

(1) \[ \text{wuz} = \text{em} \quad \text{çi} \quad \text{şew-i} = \text{m} \quad \text{şkəndevdi} \]
1SG.NOM=1SG SELF.Poss horns-ACC=1SG break.PST
‘I broke my horns.’ (Fuchs 2015:151)

(2) \[ \text{jan} \quad \text{də} \quad \text{sçıb} = \text{e} \quad \text{lup} = \text{e} \quad \text{çat-i} \quad \text{ki} \quad \text{nayardum} \quad \text{əʧ} \quad \text{waxt} \quad \text{də} \quad \text{then in old.time=IPFV big=IPFV say-PST COMP bear} \quad \text{no time with} \]
\[ \text{məltıp-əɾə} \quad \text{ja-ʃəɾib} \quad \text{mə} \quad \text{rəʈʂ-əv} \quad \text{shotgun-ABL 3SG-DAT near NEG.IMP go-2PL.SAGR} \]
‘Then in old times the old people said: Bears, never go near them (not even) with a shotgun.’ (SanGregory 2015:7).
Behind the scenes

Exploring Wakhi morphosyntax with regular expressions

- Looks for all examples with a PROGRESSIVE clitic...
  - utterance-initial: \^=PROG
  - after the first word: ^\b\w+\b =PROG
  - after the second word: ^\b\w+\b \b\w+\b =PROG
  - after the third word: ^\b\w+\b \b\w+\b \b\w+\b =PROG
  - in multiple positions: =PROG.*=PROG
Exploring Gurung phonology with regular expressions

- Looking for clusters: [pbtdkgwj][pbtdkgwj]
- Looking for geminates: (.)\1
- Looking for reduplication: (.*)(.*\*)\1\2
- Comparing frequencies of ND and NT clusters: [\mn\n][\ptk] vs. [\mn\n][\bdg]
- Looking for post-stopped nasals with following nasal vowels: [\mn\n][\bdg] ̃
- Looking for nasals with following oral vowels: [\mn\n][^:\swjrd]
An open challenge

“Apart from technical challenges, there is also an important sociological challenge to create maximally interoperable language analysis software. To imagine this can be done simply by adopting common file formats, or by operating an in-house software development lifecycle using project funds, or by invoking the XML family of buzzwords is to misunderstand the nature of the problem. Instead, we need to foster new research collaborations involving computational linguists and field linguists, leading to new understanding about how to collect and analyze corpora of data from endangered languages.”

(Bird 2009:473)
Thank you!
Bibliography


